



THE UNIVERSITY *of* EDINBURGH

## Edinburgh Research Explorer

# Phylogeography and epidemic history of hepatitis C virus genotype 4 in Africa

### Citation for published version:

Iles, JC, Raghwani, J, Harrison, GLA, Pepin, J, Djoko, CF, Tamoufe, U, LeBreton, M, Schneider, BS, Fair, JN, Tshala, FM, Kayembe, PK, Muyembe, JJ, Edidi-Basepeo, S, Wolfe, ND, Simmonds, P, Klenerman, P & Pybus, OG 2014, 'Phylogeography and epidemic history of hepatitis C virus genotype 4 in Africa', *Virology*, vol. 464-465C, pp. 233-243. <https://doi.org/10.1016/j.virol.2014.07.006>

### Digital Object Identifier (DOI):

[10.1016/j.virol.2014.07.006](https://doi.org/10.1016/j.virol.2014.07.006)

### Link:

[Link to publication record in Edinburgh Research Explorer](#)

### Document Version:

Publisher's PDF, also known as Version of record

### Published In:

Virology

### Publisher Rights Statement:

This article is available under the terms of the Creative Commons Attribution License (CC BY). You may distribute and copy the article, create extracts, abstracts, and other revised versions, adaptations or derivative works of or from an article (such as a translation), to include in a collective work (such as an anthology), to text or data mine the article, including for commercial purposes without permission from Elsevier. The original work must always be appropriately credited.

### General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

### Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [openaccess@ed.ac.uk](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.





## Phylogeography and epidemic history of hepatitis C virus genotype 4 in Africa

James C. Iles<sup>a</sup>, Jayna Raghwani<sup>a</sup>, G.L. Abby Harrison<sup>b</sup>, Jacques Pepin<sup>c</sup>, Cyrille F. Djoko<sup>d</sup>, Ubald Tamoufe<sup>e</sup>, Matthew LeBreton<sup>e</sup>, Bradley S. Schneider<sup>e</sup>, Joseph N. Fair<sup>e</sup>, Felix M. Tshala<sup>f</sup>, Patrick K. Kayembe<sup>g</sup>, Jean Jacques Muyembe<sup>h</sup>, Samuel Edidi-Basepeo<sup>i</sup>, Nathan D. Wolfe<sup>e,j</sup>, Peter Simmonds<sup>k</sup>, Paul Klenerman<sup>l</sup>, Oliver G. Pybus<sup>a,\*</sup>

<sup>a</sup> Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, UK

<sup>b</sup> Department of Infection & Immunity, Walter & Eliza Hall Institute, Victoria 3052, Australia

<sup>c</sup> Department of Microbiology and Infectious Diseases, Université de Sherbrooke, Sherbrooke, Canada

<sup>d</sup> Global Viral, Yaoundé, Cameroon

<sup>e</sup> Metabiota, San Francisco, USA

<sup>f</sup> Department of Military Health, Ministry of Defense, Kinshasa, Democratic Republic of the Congo

<sup>g</sup> Kinshasa School of Public Health, Kinshasa, Democratic Republic of the Congo

<sup>h</sup> National Institute of Biomedical Research, Kinshasa, Democratic Republic of the Congo

<sup>i</sup> National AIDS Control Program, Reference Laboratory, Kinshasa, Democratic Republic of the Congo

<sup>j</sup> Stanford University Program in Human Biology, Stanford, USA

<sup>k</sup> Roslin Institute, University of Edinburgh, Edinburgh, UK

<sup>l</sup> Peter Medawar Building for Pathogen Research, University of Oxford, Oxford, UK

### ARTICLE INFO

Available online 6 August 2014

Keywords:

Hepatitis C

Molecular epidemiology

Prevalence

Phylogeny

Molecular clock

Skyline plot

Africa

DRC

Egypt

### ABSTRACT

HCV genotype 4 is prevalent in many African countries, yet little is known about the genotype's epidemic history on the continent. We present a comprehensive study of the molecular epidemiology of genotype 4. To address the deficit of data from the Democratic Republic of the Congo (DRC) we PCR amplified 60 new HCV isolates from the DRC, resulting in 33 core- and 48 NS5B-region sequences. Our data, together with genotype 4 database sequences, were analysed using Bayesian phylogenetic approaches. We find three well-supported intra-genotypic lineages and estimate that the genotype 4 common ancestor existed around 1733 (1650–1805). We show that genotype 4 originated in central Africa and that multiple lineages have been exported to north Africa since ~1850, including subtype 4a which dominates the epidemic in Egypt. We speculate on the causes of the historical intra-continental spread of genotype 4, including population movements during World War 2.

© 2014 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/3.0/>).

### Introduction

Hepatitis C virus (HCV) is a major human pathogen that causes substantial morbidity and mortality worldwide. It is estimated that more than 185 million people are chronically infected with HCV and that there are 3–4 million new infections each year (Mohd Hanafiah et al., 2013). Infection with the virus is typically asymptomatic or unspecific in the initial stages, but once it progresses to long-term chronic infection it can lead to liver cirrhosis, fibrosis, and sometimes hepatocellular carcinoma (Lauer and Walker, 2001).

HCV is a genetically diverse virus that is classified into seven genotypes (1–7) with an average of 35% nucleotide divergence between strains belonging to different genotypes. All genotypes except 5 and 7 are subdivided into numerous subtypes (1a, 1b, 1c, 2a, 2b etc.) and the average nucleotide divergence between subtypes of the same genotype is around 25% (Murphy et al., 2007a; Simmonds et al., 1993; Smith et al., 2014).

Although the majority of HCV infections worldwide are caused by a small number of 'epidemic' subtypes (e.g. 1a, 1b, 2a and 3a) that spread rapidly during the twentieth century (Magiorkinis et al., 2009; Pybus et al., 2005; Simmonds, 2004), there are clear geographic patterns in the distribution of HCV genetic diversity. Highly divergent 'endemic' strains that belong to the same genotype are typically found in a restricted geographic area, indicating the presence of the genotype in that location for

\* Corresponding author.

E-mail address: [oliver.pybus@zoo.ox.ac.uk](mailto:oliver.pybus@zoo.ox.ac.uk) (O.G. Pybus).

hundreds or thousands of years (Simmonds, 2004). For example, HCV genotype 2 is thought to be endemic in West Africa, genotypes 1 and 4 in Central Africa and the Middle East and genotype 6 in East Asia (Candotti et al., 2003; Jeannel et al., 1998; Mellor et al., 1995; Ndjomou et al., 2003; Pybus et al., 2009).

The epidemiology of HCV prior to the discovery of the virus is poorly understood. Documentary evidence of past HCV transmission is difficult to establish as symptoms during acute infection are unspecific and HCV incidence before the widespread use of injections was likely too low to create notable outbreaks of disease. Further, samples available for retrospective screening that were archived before the 1970s are exceptionally rare (Gray et al., 2013). As a consequence, evolutionary analyses of contemporary HCV gene sequences using phylogenetic and coalescent-based methods have been utilised to estimate dates of viral divergence and to estimate viral effective population size through time. In addition, previous studies of genotype 2 in Africa (Markov et al., 2009) and of genotype 6 in Asia (Pybus et al., 2009) employed phylogeographic and molecular clock methods and provided insights into the historical geographic spread of HCV, the age of HCV genotypes and subtypes, and their recent transmission history.

To date there has been no systematic phylogeographic or evolutionary study of HCV genotype 4 as a whole. This genotype is common throughout most of Central Africa and parts of the Middle East. Recent estimates indicate that there are ~8 million people infected with HCV in Central and Eastern Sub-Saharan African, and > 15 million people infected across North Africa and the Middle East (Mohd Hanafiah et al., 2013). Genotype 4 (and subtype 4a in particular) dominates the HCV epidemic in Egypt, where 15% of adults are antibody-positive for HCV, with a much higher prevalence seen in older cohorts (El-Zanaty and Way, 2009). HCV in Egypt has been described as a 'local epidemic', whereby the transmission of one or a few subtypes rises rapidly within a region, but without the international dissemination observed for the 'global epidemic' subtypes such as 1a and 1b. High HCV seroprevalences and local epidemics associated with other subtypes of genotype 4 and have also been reported in many sub-Saharan African countries. For example, 11.2% of people screened in rural Gabon were seropositive for HCV (Njoum et al., 2012) of whom 92% were infected with genotype 4. In that study major risk factors for HCV infection were past injections, hospital admissions, and age greater than 55. In Cameroon, HCV seroprevalence was 11% in a group of high-HIV risk individuals and 16% of the HCV infections were classified as genotype 4 (Ndjomou et al., 2003). In a separate Cameroonian cohort comprising individuals aged over 60, HCV seroprevalence was 56% and 54% of infections were genotype 4 (Pépin et al., 2010a). In each of these studies HCV seroprevalence was strongly associated with age and subtypes 4a and 4r were observed.

The evolution and genetic history of genotype 4 is worthy of investigation for several reasons. First, together with genotype 1, genotype 4 responds less well to interferon-based anti-HCV drug treatment than genotypes 2 and 3, especially in patients of African descent (Chen et al., 2012; Rose et al., 2013) and it has been hypothesised that this phenotype is a consequence of the long-term presence of genotypes 1 and 4 in Central African populations (Rose et al., 2013). Second, there are a number of unanswered questions concerning the origin and spread of HCV genotype 4 within Africa. For example, the current distribution and past spread of genotype 4 strains among countries is unclear and the geographic source of the HCV lineages present in Egypt and the Middle East is currently unknown. In addition, in recent years there has been rapid growth in the prevalence of HCV subtypes 4a and 4d in Europe, particularly among injecting drug users (IDUs), hence a comprehensive overview of genotype 4 diversity may

prove useful for public health assessments outside Africa (Ciccozzi et al., 2012; De Bruijne et al., 2009; van Asten et al., 2004).

The investigation of HCV evolution in Central Africa is hampered by a lack of information about its epidemiology and genetic diversity in the region. Mohd Hanafiah et al. (2013) define the evidentiary support for HCV prevalence in the region as 'very limited'. Although recent surveillance studies have explored the genetic diversity of HCV in Cameroon, Gabon and the Republic of Congo (Cantaloube et al., 2010; Ndong-Atome et al., 2008; Nerrienet et al., 2005; Njoum et al., 2007, 2012) there is little information about the diversity of the virus in the Democratic Republic of the Congo (DRC). This country is the second largest in Africa and has 67 million residents, making it the third most populous in the continent. However, the ongoing conflict there since 1996 has made disease surveillance difficult. The large size and central position of the DRC within Central Africa mean that phylogeographic studies of HCV in the region will be incomplete without a comprehensive survey of viral diversity in the country. Further, it is possible that the DRC harbours previously-undetected variants of the virus: the only published isolate of HCV genotype 7 was isolated from a Canadian resident who had emigrated from the DRC (Murphy et al., 2007b).

At present there is little information about the genetic diversity of HCV infections in the DRC. In a previous small-scale survey of blood samples from the country (Iles et al., 2013) we detected HCV RNA in 11 individuals. Phylogenetic analysis of HCV core and NS5B region sequences from these samples indicated that they belonged to several classified and unclassified subtypes of genotype 4. In this study we address the deficit of HCV genetic information from the DRC with the screening and sequencing of 1999 blood samples from the country. We combine these new data with genotype 4 sequences gathered from online databases and originating from countries across Africa and worldwide. This enables us to analyse the DRC samples in context with the larger diversity of HCV genotype 4 viruses and to investigate the long-term evolutionary history of the virus within the African continent.

## Results

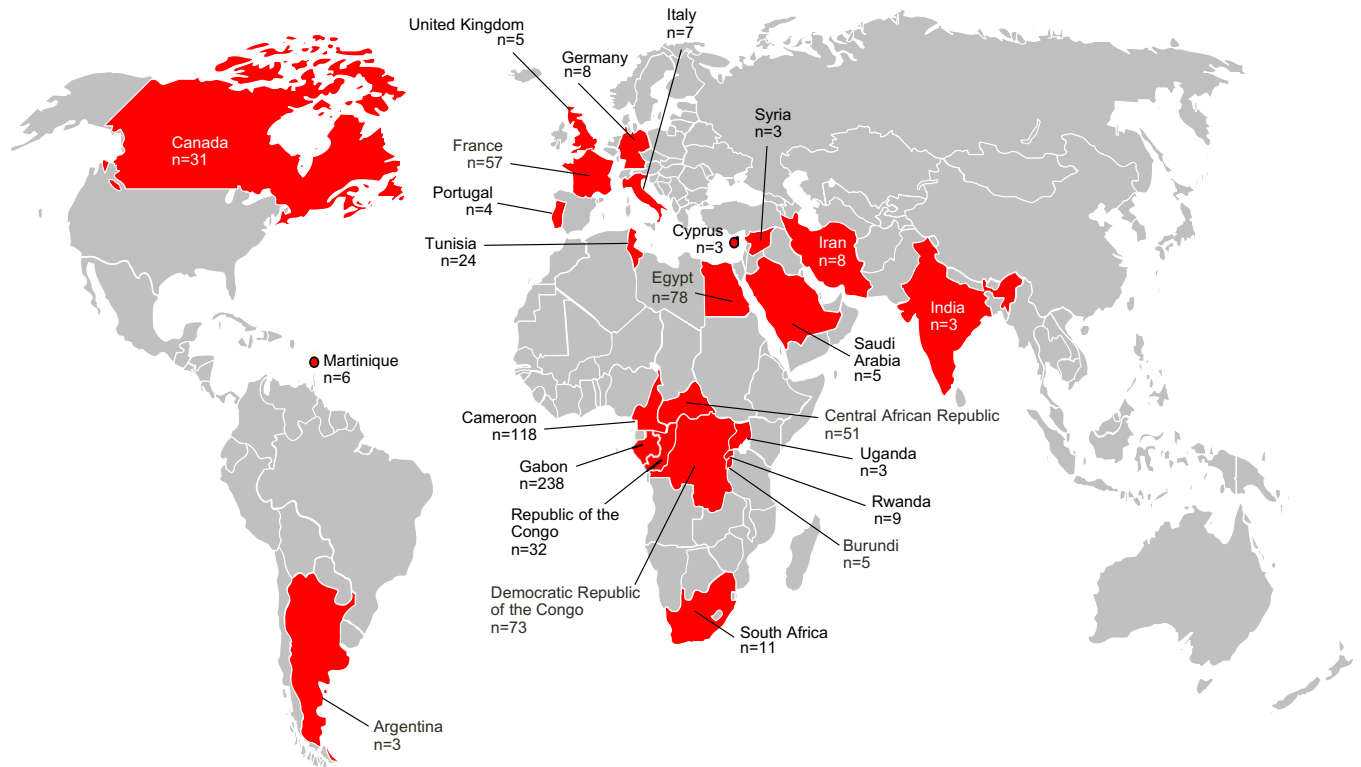
### Age distribution

Of the 1999 samples tested, 3% ( $n=60$ ) were positive for HCV 5'UTR RNA. Of these 60 samples, 33 produced core sequence and 48 produced NS5B sequence. These results are broadly consistent with our pilot study which detected HCV RNA in 3.7% ( $n=11$ ) of samples and yielded 9 core sequences and 11 NS5B sequences (Iles et al., 2013).

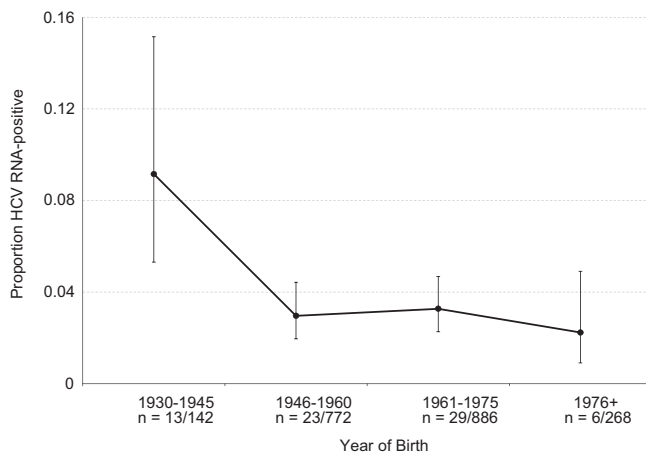
Fig. 2 shows the age distribution of HCV RNA positivity in our population, (including data from the pilot study). Samples were scored as HCV RNA positive if sequence was obtained from at least one of the 5'UTR, core or NS5B regions. When grouped according to date of birth, samples from older individuals are more likely to contain HCV RNA than younger ones; the highest prevalence (9.2%) was observed in the 1930–1945 cohort. Samples from patients born in or before 1945 were significantly more likely to contain HCV RNA than those born after ( $p < 0.001$  using Fisher's exact test).

### Phylogenetic analysis

Maximum-likelihood phylogenies estimated from the core and NS5B alignments are too large to display here and are thus provided in [Supplementary Figs. 1 and 2](#). Sequences from all formally defined subtypes of genotype 4 (Smith et al., 2014) were present in the alignments, as well as sequences from subtypes that



**Fig. 1.** Map of the world showing the country of origin and sample size of samples used in this study. Where a single isolate yielded multiple sequences, they are only counted as one sample on this figure. Countries less than three samples are not shown.



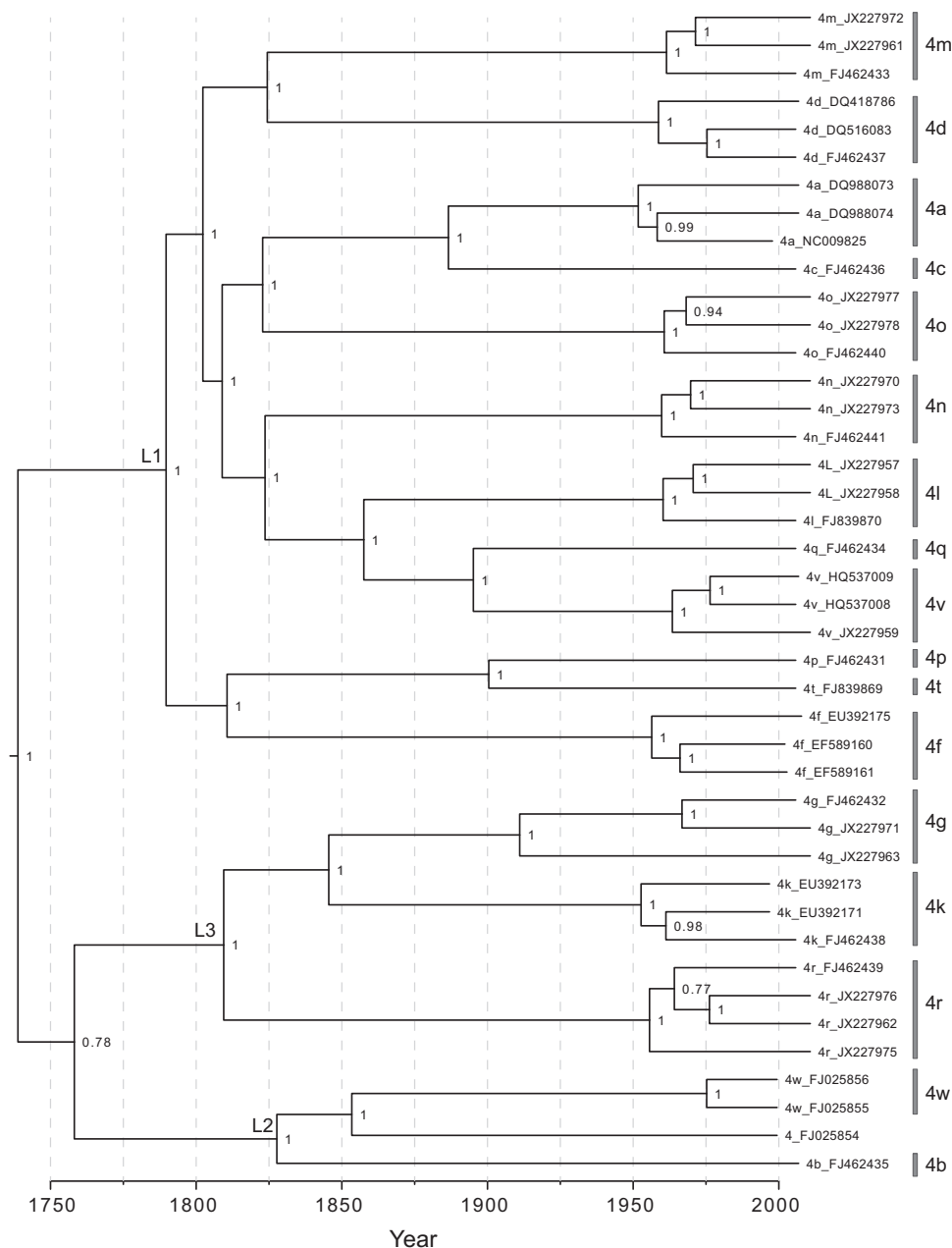
**Fig. 2.** The age distribution of HCV RNA positivity among 2298 blood samples from the DRC. Samples are assigned to one of four age categories by year of birth. Numbers below each category indicate the number of positive samples/total number of samples. Fifteen samples (all negative) did not have date of birth information and are not included. The y-axis shows the proportion of samples in each age category that were HCV RNA-positive. The errors bars represent 95% confidence limits of this proportion, estimated using the adjusted Wald method (Agresti and Coull, 1998).

are not formally defined due to a lack of whole-genome reference sequences (e.g. 4e, 4h, and 4u). The numbers of sequences assigned to each subtype and the most common locations of sampling of each subtype are shown in Table 3. A total of 74 samples were genetically too divergent to be assigned to a known subtype, 26 of which appear to belong to a provisional subtype circulating in the Central African Republic (referred to here as 4car). Sequences from five isolates (including one obtained in this

study) were discordant, i.e. they grouped into different subtypes in the core and NS5B alignments. The HCV-positive samples sequenced in this study were genetically diverse and were classified as belonging to subtypes 4c ( $n=17$ ), 4h ( $n=2$ ), 4k ( $n=18$ ), and 4r ( $n=8$ ). One sample (DRC0387) could not be classified into any known subtype but grouped with two unclassified isolates (DRC2431 and DRC2450) from our pilot study (Iles et al., 2013). This cluster of three strains is denoted 4drc here and represents a potentially new subtype (see Table 3). One sample (isolate DRC1427) generated a core sequence that was classified as 4q and a NS5B sequence classified as 4c. Few nodes in the core and NS5B maximum likelihood trees had high bootstrap support (Supplementary Figs. 1 and 2), but that is not unexpected for phylogenies estimated from these short subgenomic regions (as previously noted in Pybus et al., 2009 and elsewhere).

We discerned four clusters (denoted C1, C2, C3 and C4) that contained multiple sequences from our study population (Supplementary Figs. 1 and 2). Cluster C1 was present in subtype 4c in both the core and NS5B trees and contained 10 new DRC isolates (plus one from the pilot study). Cluster C2 was also found in subtype 4c and comprised six isolates from this study in the NS5B tree (only four of which are present in the core tree). Cluster C3 contained 12 subtype 4k samples from this study (plus two from the pilot study). C3 also included 15 samples from Tunisia in the NS5B tree, whereas in the core tree it included seven sequences from Gabon. Finally, cluster C4 was present in subtype 4r. In the NS5B tree C4 contained eight sequences from this study (plus two from the pilot study).

Samples from our study population also appeared outside of these four clusters. Specifically, two further isolates were placed inside subtype 4c, nine within 4k, two within 4h, one within 4q and three grouped in the unclassified lineage 4drc (see above). The pair of subtype 4h sequences grouped significantly, together with a strain from Brazzaville (Republic of the Congo; GU088141).



**Fig. 3.** Maximum clade credibility molecular clock phylogeny, estimated from the whole genome alignment. Branch lengths represent time (see scale bar at the bottom of the figure). Posterior clade probabilities are shown next to each node. Sequences are labelled with their subtype and accession number. Subtypes are indicated on the right side of the diagram. The three intra-genotypic lineages discussed in the main text are labelled L1, L2, and L3.

### Whole genome phylogenies

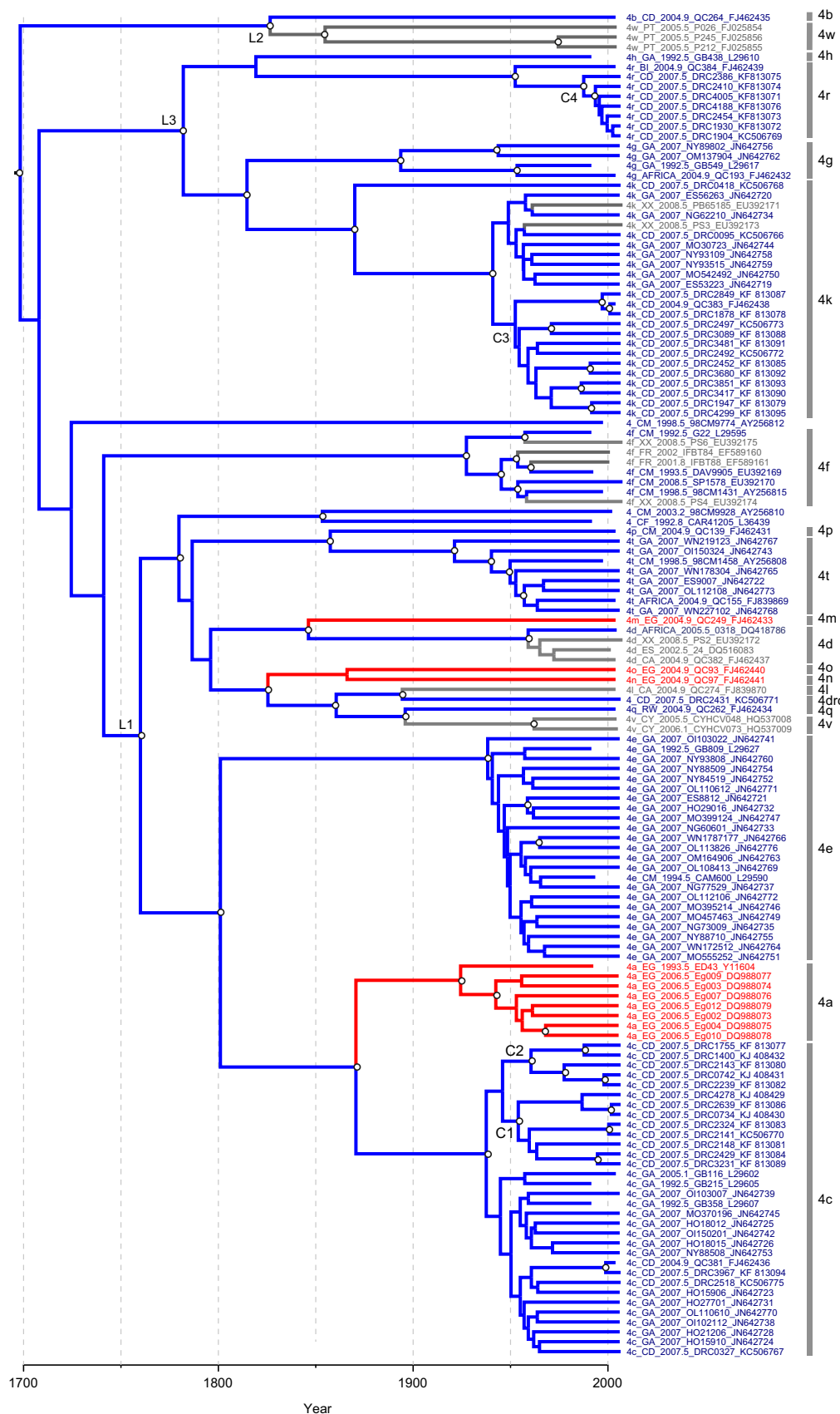
Fig. 3 shows a maximum clade credibility (MCC) phylogeny of HCV genotype 4, obtained from the Bayesian molecular clock analysis of the whole genome sequences. An equivalent maximum likelihood phylogeny is provided in [Supplementary Fig. 3](#). Up to three genomes per subtype were included and no complete genome sequences were available for some subtypes (subtypes 4e, 4h, 4s and 4u). As Fig. 3 indicates, there is a great deal of phylogenetic structure above the subtype level. This structure is supported by high posterior probability values in the Bayesian phylogeny and by high bootstrap values in the ML phylogeny (Fig. 3; [Supplementary Fig. 3](#)).

Three intra-genotypic lineages can be discerned and are denoted here L1, L2, and L3. These lineages correspond to the three clades closest to the root of genotype 4 with strong statistical

support. Lineage L1 contains subtypes 4a, 4c, 4d, 4l, 4m, 4n, 4o, 4q and 4v, L2 contains subtypes 4b and 4w and L3 contains subtypes 4g, 4k, and 4r. In the ML phylogeny, L2 is present as an outgroup ([Supplementary Fig. 3](#)) whereas in the Bayesian tree it is placed as a sister group to L3, albeit with a comparatively low posterior probability of 0.78 (Fig. 3).

The Bayesian molecular clock analysis provided an estimate of the date of the most recent common ancestor (MRCA) of genotype 4, which was 1733 (95% HPD credible region = 1650–1805). This date is more recent than some previous estimates for the origin of genotype 4. [Pybus et al. \(2001\)](#) estimated HCV evolutionary rates from a small data set of dated sequences and used these to infer that the MRCA of genotype 4 existed about 350 years ago. [Njoum et al. \(2007\)](#) used the same rate estimates during a more comprehensive Bayesian phylogenetic analysis and dated the MRCA of genotype 4 to 1541 (95% CIs: 1343–1698). The more recent date estimated here is likely





**Fig. 4.** Maximum clade credibility molecular clock phylogeny, estimated from the concatenated alignment. Taxa are coloured according to location of sequence origin (blue=sub-Saharan Africa; red=Middle East and North Africa; grey=rest of the world). The locations of internal branches were inferred using parsimony and are coloured similarly. Branch lengths represent time (see scale bar at the bottom of the figure). Nodes with a posterior probability > 0.9 are labelled with a white circle. Sequences are labelled as follows: subtype, sampling location using two-letter country codes (ISO 3166; see Table 4), sampling date, isolate name, accession number. XX represents an unknown location. Subtypes are indicated on the right side of the diagram. The three intra-genotypic lineages discussed in the main text are labelled L1, L2, and L3. The four clusters of samples obtained in this study discussed in the main text are labelled C1, C2, C3, and C4.

to be more accurate because (i) it is based on whole genome sequences rather than small subgenomic fragments and (ii) it employs new HCV evolutionary rates that were estimated using larger data sets of dated sequences and more powerful methods of analysis (Gray et al., 2011). We note that the MRCA date provided here could be underestimated as a result of strong purifying selection (Wertheim and Kosakovsky Pond, 2011) or overestimated due to strong among-branch rate variation (Wertheim et al., 2012). However, Markov et al. (2012) showed that molecular clock estimates of HCV lineage movement between Africa and the Americas matched the known timeframe of the trans-Atlantic slave trade. This suggests that among-subtype HCV divergence dates within a genotype can be estimated with reasonable accuracy.

#### Concatenated alignment phylogenies

Fig. 4 shows the MCC phylogeny obtained from the Bayesian molecular clock analysis of the concatenated sequences (core plus NS5B). As in the whole genome phylogenies, all subtypes were monophyletic, and three well-supported intra-genotypic lineages were again observed. The greater number of subtypes present in concatenated analysis indicates that L1 contains subtypes 4a, 4c, 4d, 4e, 4l, 4m, 4n, 4o, 4p, 4q, 4t and 4v, L2 contains subtypes 4b and 4w, and lineage L3 contains subtypes 4g, 4h, 4k and 4r. L2 is again placed in a basal position (as in the full-genome ML phylogeny; Supplementary Fig. 3). However the weak posterior probability for this placement (0.21) suggests that L2 may be in reality be a sister group to L3 (as in the full genome Bayesian phylogeny; Fig. 3). Subtype 4f plus one divergent isolate from Cameroon (98CM9774) did not group into any of these three well supported intra-genotypic lineages.

The concatenated alignment yielded an estimate of the date of the genotype 4 common ancestor of 1687 (95% HPD credible region: 1538, 1811). This estimate overlaps considerably with the estimate from the whole genome alignment and the upper 95% HPD limits of both are almost identical. The credible region for the concatenated alignment is wider than that for the whole genome alignment, likely reflecting the reduced phylogenetic information in the former.

The molecular clock analysis was also used to estimate the date of origin of the four sequence clusters (C1–4) that contained isolates from our study population (see above). The estimated date of the MRCA of cluster C1 was 1953 (95% HPD: 1938, 1982). For cluster C2 the MRCA was dated to 1960 (95% HPD: 1930, 1976) and for cluster C3 it was 1952 (95% HPD: 1928, 1975). The estimated MRCA for cluster C4 was somewhat more recent and was dated to 1987 (1976, 2000). While these clusters had low bootstrap scores in the ML trees (Supplementary Figs. 1 and 2), in this analysis C1, C2 and C4 are all supported with posterior probabilities > 0.9, while C3 has a posterior probability of 0.79. This increased statistical support is likely due to the combined phylogenetic signal in the concatenated analysis as compared to that available in the ML reconstruction of individual genome regions.

The taxa and branches of Fig. 4 have been coloured according to the known or inferred country of origin for each isolate. The majority of strains are from sub-Saharan Africa (blue). Sequences from North Africa/Middle East (red) and from the rest of the world (grey) cluster together within the greater diversity of genotype 4 from sub-Saharan Africa. Thus it is clear that genotype 4 originated in central Africa before disseminating elsewhere. Further, some isolates without location information or which were sampled outside Africa (e.g. subtypes 4f and 4k strains closely related to those from Cameroon and Gabon) may also represent infections that were acquired in Central Africa. Almost all isolates sampled outside central Africa belong to the intra-genotypic lineage L1 (Fig. 4). Sequences from Egypt are commonly found in subtypes 4a, 4l, 4m, 4n and 4u (Supplementary Figs. 1 and 2;

Table 3). Subtype 4d is currently found in many countries, especially in Western Europe, where it is prevalent among some injecting drug user populations (De Bruijne et al., 2009).

Fig. 4 provides some clues as to the origin of the most common subtypes of genotype 4. For example, subtype 4a is found worldwide but is highly prevalent in Egypt, where it was likely spread by widespread injection treatment campaigns during the mid-twentieth century (e.g. Strickland, 2006; Pybus et al., 2003). Using our results we can explore the question of from where the Egyptian HCV epidemic originated. Our analysis indicates that subtypes 4a and 4c diverged from each other around 1870 and that subtype 4c mainly comprises sequences from Gabon and the DRC. In the core and NS5B trees (Supplementary Figs. 1 and 2) we can observe strains sampled from Cameroon and the Central African Republic that are phylogenetically immediately basal to subtypes 4a, 4n and 4o. Fig. 4 indicates that these subtypes (and subtype 4m) arrived in Egypt from Central Africa no earlier than 1825.

Fifteen subtype 4k isolates from Tunisia cluster together within the much greater diversity of subtype 4k sampled from central Africa. The Tunisian strains are closely related to DRC isolates from this study belonging to cluster C3. Although these Tunisian samples were not included in the concatenated analysis (because only NS5B sequences were available) we can combine Supplementary Figs. 2 and 4 to conclude that the Tunisian subtype 4k infections originated from central Africa (possibly from the DRC) at around the time of the MRCA of cluster C3, which was 1952 (1928, 1975). A second cluster of six unclassified Tunisian isolates can be observed in the NS5B phylogeny and form a sister lineage to subtype 4o (Supplementary Fig. 2).

#### Bayesian skyline analysis

The epidemic history of genotype 4 was investigated by estimating a Bayesian skyline plot from the concatenated alignment (Fig. 5). This plot represents the effective number of HCV infections through time, back to the estimated TMRCA of the genotype. Fig. 5 shows that prior to the twentieth century there was an extended period of endemic transmission of genotype 4, during which the effective number of infections was low. There is a sharp transition from endemic infection to rapid exponential growth starting around the 1950s. The epidemic history of genotype 4 after 1975 is much harder to discern due to very large confidence limits: either scenario after 1975 of continued growth or a stabilisation of prevalence are statistically compatible with the data. Previous studies of genotype 4 in several African countries (Central African

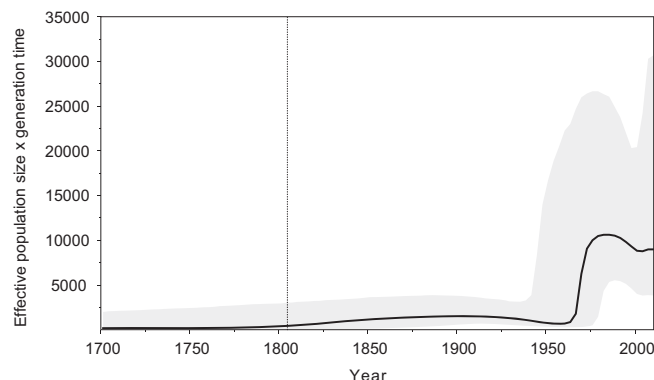


Fig. 5. Bayesian skyline plot, showing the epidemic history of genotype 4 estimated from the concatenated alignment. The black line represents the estimated effective number of infections through time. The blue lines represent the 95% highest posterior density confidence intervals of this estimate. The earliest date in the plot is the median estimate of the TMRCA of genotype 4, while the dotted line shows the upper 95% highest posterior density confidence interval of this date.

Republic, Republic of Congo, Gabon, Cameroon) have noted similar, but earlier, transitions to rapid expansion, between 1930 and 1960 (Cantaloube et al., 2010; Njouom et al., 2007, 2009, 2012). The slightly later date of the epidemic transition reported here is likely explained by the new evolutionary rates used in our analysis, which are faster than those employed previously (see above for discussion). The overall shape of the epidemic curve presented in Fig. 5 closely resembles that previously estimated for genotype 6 in Asia (Pybus et al., 2009). However, as discussed in that paper, skyline plots that span entire HCV genotypes should be interpreted cautiously, as they exhibit geographic structure and combine subtypes that have experienced different rates of growth during the twentieth century (see Pybus et al., 2009 for further details). Thus only the broad qualitative trend of a twentieth-century transition to epidemic growth is likely to be robust.

## Discussion

In this study we report the first large-scale survey of HCV genetic diversity in the DRC, one of the geographically largest and most populous African countries. The previously unsurveyed state of the DRC represented a significant gap in our understanding of HCV in Africa and our results complement recent surveys of HCV in neighbouring countries, including Gabon, Cameroon, Republic of Congo and Central African Republic (e.g. Cantaloube et al., 2010, Njouom et al., 2007, 2009, 2012, Pépin et al., 2008, 2010a, 2010b). We observed a high level of HCV genotype 4 diversity in our study population; isolates were classified as belonging to subtypes 4c, 4k, 4h and 4r, as well as to a potential new subtype circulating in the DRC (denoted 4drc). One isolate (DRC1427) was classified discordantly (4c/4q) in two sub-genomic regions sequenced and may represent a previously undetected recombinant strain. This diversity, together with the absence of globally-prevalent subtypes (e.g. 1a, 1b, and 3a), suggests that genotype 4 has been present in the DRC for a long period of time. However, the genetic diversity of HCV in the DRC is not as high as that in other central African countries where various subtypes of multiple HCV genotypes are sometimes observed (e.g. genotypes 1, 2 and 4 in Cameroon; Ndjomou et al., 2003, Pépin et al., 2010a). Further, we failed to detect any genotype 7 isolates despite using the same primers as those used to initially discover the strain (Murphy et al., 2007b), indicating that the prevalence of genotype 7 within the DRC is low. We conclude that the overall pattern of genetic diversity of HCV in the DRC is fundamentally different to that of HIV-1; the diversity of HIV-1 in the DRC is as large as that observed worldwide, implying that the virus originated in central Africa and its diversity elsewhere is reduced due to founder effects (Vidal et al., 2000; Rambaut et al., 2001), whereas we observed only one HCV genotype in the DRC.

Approximately 3% of our samples screened contained detectable HCV RNA, slightly less than the 5.2% and 6.4% found in HCV screening surveys in Gabon and Central African Republic respectively (Njouom et al., 2009, 2012). Past studies of HCV in the DRC have only assayed seroprevalence, and have found anti-HCV in 6.4% of blood donors (Tibbs et al., 1991), 6.6% of sex workers and 4.3% of pregnant women (Laurent et al., 2001). These seroprevalence estimates are again somewhat less than those found in the Central African Republic (10.5%) and Gabon (11.2%) (Njouom et al., 2009, 2012). Despite the large size of our survey, sampling was limited to males and therefore our prevalence estimates may not equal those of the general population. Cantaloube et al. (2010) found that men were 43% more likely to be seropositive than women in the Republic of the Congo, and it is possible that a similar pattern is true in the DRC. We confirmed that HCV

positivity is significantly associated with older age in our study population (Fig. 2; Iles et al., 2013). Although the long-term persistence of most HCV infections means that older subjects are in general more likely to be infected, the notable increase in HCV prevalence in those born before 1945 suggests a distinct historical event. This 'cohort effect' has also been reported for HCV genotype 4 in Gabon, Cameroon and Egypt and is thought to be due to past iatrogenic transmission resulting from public health campaigns during the twentieth century that involved injections (Frank et al., 2000; Nerrienet et al., 2005; Njouom et al., 2007, 2012). In the absence of any epidemiological data we cannot speculate further on the possible causes in our study population of the age effect. However, our reconstruction of the epidemic history of genotype 4 (Fig. 5) clearly shows a transition from low to high prevalence, likely representing the combined effects of iatrogenic transmission events in many different African countries during the middle of the twentieth century (Pépin and Labbé, 2008).

The molecular clock and phylogeographic results reported here clearly indicate that genotype 4 epidemics in North Africa, the Middle East and Europe all originated from central Africa (Fig. 4). Fig. 4 can be also used to estimate the time of exportation of lineages from central Africa to elsewhere; however, the ages of these movement events will often be overestimated due to the exclusion from the concatenated alignment of many central African strains that were sequenced in only the core or NS5B regions (Supplementary Figs. 1 and 2). Genotype 4 lineages present in Egypt are of particular interest, as the country has the highest prevalence of HCV worldwide (Frank et al., 2000; El-Zanaty and Way, 2009). We estimate that subtype 4a, which accounts for the majority of the Egyptian epidemic, arrived there from central Africa sometime between 1860 and 1925 (the estimated date of the MRCA of subtype 4a). This time frame precedes World War II, during which an 8000-man contingent of the Force Publique (an army composed of soldiers from the Congo serving in the Free Belgian Forces) was stationed in Egypt between 1943 and 1944 (Nigel, 1991). However, these dates can be reconciled if two or more genetically-distinct subtype 4a lineages arrived in Egypt during World War II. It is clear that the Egyptian HCV epidemic originated from more than one Central Africa strain, as subtypes 4a, 4m, 4n and 4o each independently moved to Egypt. Stronger evidence for the role of historical events in the trans-African movement of genotype 4 comes from the presence of Tunisian isolates closely related to strains from the DRC, within subtype 4k. Our molecular clock results date the MRCA of the Tunisian 4k strains to between 1928 and 1975. Tunisian troops were deployed in the DRC as part of the UN peacekeeping forces during the Congo crisis of 1960–65 (Mays, 2010) and may therefore have transported the virus to Tunisia upon their return. The movements of large numbers of troops among populations, combined with a high likelihood of blood transfusions and/or parenteral medical treatments, provided an ideal scenario for the spread of HCV out of central Africa.

Our NS5B phylogeny contains a very high diversity of genotype 4 sequences sampled from France (Supplementary Fig. 2). This can be explained by the historical presence of the French colonial empire in central Africa, followed by more recent immigration of individuals to France from former colonies (see also Nicot et al., 2005). We also noted that subtype 4w has, to date, been found only in Portugal. Portugal was one of the first nations to have a foothold in the Congo region (Appiah and Gates, 1999) and was present in both mainland Angola and in the Cabinda exclave that lies between the DRC and the Republic of the Congo. Our phylogenetic results suggest that subtype 4w is a sub-lineage of 4b, the latter being a highly diverse subtype found in several countries, including countries that neighbour Cabinda (Fig. 3; Supplementary Fig. 2).



Other previously unrecognised geographic trends can be discerned from our analysis. For example, the majority of isolates belonging to subtypes 4q and 4v (9 out of 14) were sampled from Rwanda and Burundi, and samples from these two countries are rarely found in other subtypes. This implies that there has been little further viral dissemination from these countries following the introduction of subtypes 4q and 4v there. The reverse pattern is observed in several adjacent countries of central Africa, specifically the DRC, Republic of the Congo, Cameroon and Gabon. Multiple subtypes of genotype 4 are present in each of these countries, and sequences from them appear in all three of the intra-genotypic lineages L1, L2 and L3. This indicates that historical viral movement among these locations was comparatively common. This might be explained by historical demographic movements and trading links during the African colonial era. However, genotype 4 does not appear to have spread to the Americas via the slave trade, unlike genotype 2 which is endemic in west Africa (Markov et al., 2009). This may reflect the various ways that European colonial powers exploited different African regions; the western Atlantic coast, including the Gold Coast, was the main hub for the transatlantic slave trade, whereas Central and Eastern Africa were more commonly used to provide natural resources and labour (Suret-Canale, 1971).

Approximately 9% (74 out of 806) of all isolates examined in our study are not classified into a currently defined subtype, and all unclassified isolates were from central African countries. Detection of divergent yet uncommon strains is to be expected if HCV genotype 4 originated and has circulated endemically in central Africa for several centuries. Under this hypothesis the more prevalent subtypes of genotype 4 likely represent those few lineages that by chance were amplified by changes in human behaviour during the previous century (as previously suggested for subtypes 1a, 1b, 3a globally, and for genotype 6 in Asia; e.g. Simmonds, 2004; Magiorkinis et al., 2009; Pybus et al., 2009). We noted two lineages (4car and 4drc) that represent candidates for possible new subtypes. Full genome sequencing of isolates from 4car and 4drc would confirm their subtype status. Lineages that are currently rare and geographically-restricted could potentially spread epidemically and internationally if introduced into high risk groups, as presumably occurred to subtype 4d (De Bruijne et al., 2009; Ciccozzi et al., 2012). This risk is of particular relevance to genotype 4 which, like genotype 1, is more refractory to standard anti-viral drug treatment than other genotypes (e.g. Chen et al., 2012). There are no large scale surveys of HCV diversity in several highly populous African countries—Ethiopia, Tanzania, Angola and Chad among them—and some other countries are represented only by surveys with small sample sizes. Further screening of HCV genetic diversity in Africa is required to help plan effective treatment strategies in the region and inform future vaccine design.

## Methods

### Study population

A total of 1999 EDTA blood samples were collected from informed consenting members of the uniformed services as part of a screening programme for infectious diseases. This collection has been studied previously for HIV-1 (Djoko et al., 2011), human parvovirus 4 (Sharp et al., 2010) and human pegiviruses (Iles et al., 2013). A preliminary small scale survey ( $n=299$ ) of this collection for HCV discovered HCV RNA in ~4% of samples (Iles et al., 2013). Collection took place during 2007 in Kinshasa, capital of the DRC. The samples were anonymised although patient year of birth were available for most. All samples were from male individuals, whose mean age was 44 (range 22–77 years).

### Screening, RT-PCR and sequencing

All samples were tested for HCV RNA. Viral RNA was extracted from sera using the Nucleospin 96 RNA kit (Macherey-Nagel) as per the manufacturer's instructions. The reaction product was screened for HCV RNA with a one-step RT-PCR amplification of the 5'UTR region using Superscript III with Platinum *taq* (Invitrogen, Life Sciences); primers are listed in Table 1. Samples positive for HCV RNA were subsequently amplified and sequenced in the core and NS5B regions using the same enzymes as used for the 5'UTR, noted above. Controls were run in parallel at each step. Primers were obtained from previous studies or were designed using a large alignment of whole HCV genome sequences that included subtypes belonging to all 7 genotypes (Table 1). The internal primers were used for sequencing with BigDye Terminator (Applied Biosystems). Traces were examined using 4Peaks (Nucleobytes). A total of 34 core sequences (accession numbers KF813071–KF813095, KJ408429–KJ408436 and KJ416140) and 48 NS5B sequences (KF826150–KF826197) were obtained in this study.

### Sequence collation

All available genotype 4 sequences were downloaded from the Los Alamos HCV sequence database (Kuiken et al., 2005) and from GenBank. Sequences were retained if they spanned either of the two subgenomic regions sequenced in this study: core (positions 342–1265 relative to H77) or NS5B (positions 8265–8624). Only one sample per region from each infected individual was retained and sequences from non-human subjects were excluded, as were sequence fragments shorter than 200 nucleotides. We noted a disproportionate number of sequences from subtype 4a, largely resulting from the high number of published studies concerning the Egyptian HCV epidemic. To bring the number of 4a sequences approximately in line with those of other subtypes we randomly

**Table 1**  
Details of primers used in this study.

Primer name	Source	Sequence (5'–3')	Position <sup>a</sup>
Murphy 5'UTR F	Murphy et al. (2007b)	GAAAGCGTCTAGCCATGGCGTTAGT	71–95
Murphy 5'UTR R	Murphy et al. (2007b)	CTCGAAGCACCCTATCAGG	311–292
5' UTR Ex 400F	This study	CCTGTGGTACTGCCTGATAG	279–299
CHV core 980 Rex	This study	AGTGCCARRAGGAACATAGA	883–864
5'UTR In 405 F	This study	CTGATAGGGTGCTTGCAGTG	293–313
CHV core 973 Rin	This study	AGTGCCARRAGGAAGATAGARAA	883–861
NS5B Ex 8274 F	This study	TGGGGATCCCGTATGATACCCGCTGCTTGA	8245–8275
NS5B Ex 8616 R	This study	CGGAATTCCTGGTCATAGCCTCCGTGAA	8643–8616
NS5B In 8378 for	This study	GACACCCGCTGCTTTGACTC	8259–8278
NS5B In 8611 rev	This study	GAGTCTTCACGGAGGCTATGACNAGGTA	8638–8611

<sup>a</sup> Numbering relative to isolate H77 (Genbank accession number AF009606).

removed sequences separately from each of the three main sub-genotypic lineages of the 4a phylogeny, thereby maintaining the full genetic diversity of the subtype in our data set. We also reduced in a similar manner the disproportionate number of subtype 4d sequences sampled in France.

The reference database sequences were collated with the new sequences obtained in this study, resulting in a total of 806 isolates across all genome regions. All sequences were aligned by hand using Se-Align v2.0 (available from <http://tree.bio.ed.ac.uk>), resulting in a 'core alignment' containing 177 core sequences and an 'NS5B alignment' containing 765 NS5B sequences. Subsequently, a 'concatenated alignment' was created by combining and concatenating core and NS5B sequences if they were sampled from the same individual and covered both genome regions. The resulting joint alignment contained 136 taxa. In order to best estimate the branching order among HCV subtypes within genotype 4, we also compiled a 'whole genome' alignment that contained all genotype 4 reference genome sequences described in Smith et al. (2014).

For each isolate we surveyed online databases and the primary literature for two pieces of information: year of sampling and country of origin. Most isolates (83%) were sampled from African or Middle Eastern countries. A search of the primary literature revealed that some HCV strains sampled in Europe or North America represent infections from recent immigrants from Africa or the Middle East. In these instances, the 'country of origin' of the infection is defined as the country from which the individual emigrated. A summary of the geographic distribution of the sequences used in this study is provided in Fig. 1.

#### Phylogenetic analysis

Phylogenies were estimated for the core, NS5B and whole genome alignments using maximum likelihood (ML) as implemented in GARLI v0.951 (Zwickl, 2006). The analysis used a General Time-Reversible (GTR) nucleotide substitution model, estimated base frequencies, and a gamma distribution model of among-site rate variation. Statistical support for phylogenetic clustering was calculated using an ML bootstrap approach with 500 bootstrap replicates; bootstrap scores were summarised using TreeAnnotator (<http://beast.bio.ed.ac.uk/TreeAnnotator>). Phylogenies were visualised and annotated using FigTree v1.4 (<http://tree.bio.ed.ac.uk/software/figtree>). Newly-generated sequences were classified by computing *p*-distances to the HCV subtype reference sequences provided in Smith et al. (2014). A *p*-distance threshold of <0.1 was used to assign subtypes. *p*-distances between sequences were calculated using DNAdist in the Phylip package (Felsenstein, 1989).

#### Calibration of the molecular clock

Molecular clock models can be used to reconstruct the evolutionary history of HCV genotype 4 on a timescale of years, provided that estimates of the rate of molecular evolution are available for the viral genomes regions being investigated. As in previous studies, we cannot directly estimate reliable HCV evolutionary rates from the alignments

in hand as the range of sampling times is too narrow (Pybus et al., 2009; Salemi and Vandamme, 2002). Therefore we estimated evolutionary rates for the core and NS5B regions using independent sets of HCV sequences that have been shown to contain good temporal information. These estimates were then used as informative prior distributions for evolutionary rate parameters in all subsequent Bayesian evolutionary analyses (see next section).

Gray et al. (2011) undertook a comprehensive analysis of HCV evolutionary rates and we use their alignments to estimate rates that are specific to the core and NS5B genome regions sequenced here (positions 342–945 and 8265–8624, respectively). We analysed two alignments, comprising 65 subtype 1a sequences and 54 subtype 1b sequences, respectively (Gray et al., 2011). These rates are likely to be accurate for our genotype 4 study because genotypes 1 and 4 are more closely related than other subtypes (Salemi and Vandamme, 2002) and because HCV evolutionary rates vary considerably more between genome regions than they do between genotypes and subtypes (Gray et al., 2011). Evolutionary rates were estimated using the Bayesian Markov Chain Monte Carlo (MCMC) inference method implemented in BEAST v1.7.5 (Drummond et al., 2012). These analyses employed a SDR06 nucleotide substitution model (two independent HKY+Γ substitution models—one for the first and second codon positions, and one for the third), an uncorrelated lognormal relaxed molecular clock model, and a Bayesian skyline plot coalescent model (Shapiro et al., 2006). Nucleotide frequencies were estimated from the data. Rates of molecular evolution were estimated for three different partitions of the whole genome alignment: (i) a core partition (sites 342–945), (ii) a NS5B partition (sites 8265–8624),

**Table 3**  
Subtype summary.

Subtype	Samples (Core)	Samples (NS5B)	Most common sampling location of core and NS5B sequences
4a	11	46	Egypt
4b	1	10	Democratic Republic of the Congo
4c	35	80	Gabon
4d	17	45	France
4e	29	142	Gabon
4f	16	106	Cameroon
4g	4	7	Gabon
4h	1	15	Cameroon
4k	23	77	Democratic Republic of the Congo
4l	1	10	Egypt
4m	2	14	Egypt
4n	1	7	Egypt
4o	3	24	Egypt
4p	1	13	Cameroon
4q	2	6	Rwanda
4r	12	40	Democratic Republic of the Congo
4t	9	31	Cameroon
4u	1	14	Egypt
4v	2	7	Rwanda
4w	5	3	Portugal
4car	0	26	Central African Republic
4drc	0	3	Democratic Republic of the Congo
Unclassified	0	45	

**Table 2**  
Estimated evolutionary rate parameters.

Genome region	Genome positions <sup>a</sup>	Estimated nucleotide substitution rate (subs/site/year)	95% credible region
(i) Core	342–944	$5.39 \times 10^{-4}$	$3.41–7.46 \times 10^{-4}$
(ii) NS5B	8274–8612	$9.87 \times 10^{-4}$	$6.74–14.4 \times 10^{-4}$
(iii) Concatenated (Core+NS5B)	342–944 and 8274–8612	$7.43 \times 10^{-4}$	$4.91–10.4 \times 10^{-4}$
(iv) Complete genome	342–9374	$13.5 \times 10^{-4}$	$9.97–17.0 \times 10^{-4}$

<sup>a</sup> Numbering relative to isolate H77 (Genbank accession number AF009606).

**Table 4**  
Two-letter country codes used in this study.

Code	Country name
AR	Argentina
BE	Belgium
BH	Bahrain
BI	Burundi
CA	Canada
CD	Democratic Republic of the Congo
CF	Central African Republic
CG	Republic of the Congo
CM	Cameroon
CY	Cyprus
DE	Germany
DK	Denmark
DZ	Algeria
EG	Egypt
ES	Spain
FR	France
GA	Gabon
GB	United Kingdom
IN	India
IR	Ireland
IT	Italy
JP	Japan
MQ	Martinique
NL	Netherlands
PK	Pakistan
PT	Portugal
RU	Russia
RW	Rwanda
SA	Saudi Arabia
SY	Syria
TN	Tunisia
UG	Uganda
US	United States of America
XX	Unknown
YE	Yemen
ZA	South Africa

and (iii) a concatenated core+NS5B partition (sites 342–945 plus 8265–8624). The rate parameters estimated for these three partitions are shown in Table 2. Each MCMC analysis was run for at least 100,000,000 states. Table 2 also includes evolutionary rate estimates for whole genome sequences, which were taken directly from Gray et al. (2011).

Bayesian evolutionary analysis

To estimate the epidemic and movement history of HCV genotype 4 we analysed the ‘whole genome’ and ‘concatenated’ alignments using the Bayesian Markov Chain Monte Carlo (MCMC) inference method implemented in BEAST v. 1.7.5 (Drummond et al., 2012).

As with the evolutionary rate estimation analyses described above, we used the SDR06 substitution model, an uncorrelated lognormal relaxed molecular clock, and a Bayesian Skyline coalescent model with 10 groups. For both the ‘whole genome’ and ‘concatenated’ data sets, Bayesian model selection tests showed that the SDR06 substitution model substantially outperformed the GTR+ $\Gamma$  model (Bayes Factor > 100; calculated using Tracer v1.5). For both data sets, a normal prior distribution was placed on the mean evolutionary rate parameter, such that the mean and variance of the prior distribution matched the ‘concatenated’ and ‘whole genome’ rate estimates shown in Table 2. Each MCMC run contained 200 million states, sampled once every 5000 states; trees were sampled every 50,000 states. Multiple MCMC runs were calculated to ensure convergence and were combined to increase the accuracy of parameter estimates. MCMC convergence and effective sample sizes were monitored using Tracer v. 1.5. Maximum clade credibility trees were calculated and annotated using

TreeAnnotator 1.7.5 (Drummond et al., 2012). FigTree v1.3.1 was used to colour lineages according to their sampling location using the parsimony criterion. We deliberately chose not to apply more sophisticated Bayesian discrete state phylogeographic models (e.g. Lemey et al., 2009) to our data. Such models are highly parametric and are unlikely to be informative when applied to phylogenies with comparatively few location state changes and no sampled sequences close to the phylogeny root (such as the tree presented in Fig. 4).

Acknowledgments

We thank the Ministry of Public Health and the Ministry of National Defense of the Democratic Republic of Congo for permission to undertake this study and the U.S. Embassy of the DRC for their continued support. This project was financially supported by Global Viral and Metabiota via funding from the U.S. Department of Defense Armed Forces Health Surveillance Center, Division of Global Emerging Infections, Surveillance Operations (AFHSC GEIS), the Defense Threat Reduction Agency Cooperative Biological Engagement Program (DTRA-CBEP), Google.org, the Skoll Foundation, the DoD HIV/AIDS Prevention Program (DHAPP), and was made possible by the generous support of the American people through the United States Agency for International Development (USAID) Emerging Pandemic Threats PREDICT. P.K. is an NIHR Senior Investigator and he and GLAH were supported by The Wellcome Trust (WT 091663MA), Oxford Martin School, and NIHR Biomedical Research Centre, Oxford. C.F.D. was supported in part by the NIH Fogarty International Center AIDS International Training and Research Program (2 D 43 TW000010-16/17). N.D.W. is supported by the NIH Directors Pioneer Award (DP1-OD000370). The contents are the responsibility of the authors and do not necessarily reflect the views of the DoD, USAID, or the United States Government.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at <http://dx.doi.org/10.1016/j.virol.2014.07.006>.

References

Agresti, A., Coull, B.A., 1998. Approximate is better than “Exact” for interval estimation of binomial proportions. *Am. Stat.* 52, 19–126.

Appiah, K.W., Gates, H.L., 1999. *Africana: The Encyclopedia of the African and African American Experience* Running Press, Philadelphia, PA p. 1105.

Candotti, D., Temple, J., Sarkodie, F., Allain, J.P., 2003. Frequent recovery and broad genotype 2 diversity characterize hepatitis C virus infection in Ghana, West Africa. *J. Virol.* 77, 7914–7923.

Cantaloube, J.-F., Gallian, P., Bokilo, A., Jordier, F., Biagini, P., Attoui, H., Chiaroni, J., de Micco, P., 2010. Analysis of hepatitis C virus strains circulating in Republic of the Congo. *J. Med. Virol.* 82, 562–567.

Chen, Y., Xu, H.X., Wang, L.J., Liu, X.X., Mahato, R.I., Zhao, Y.R., 2012. Meta-analysis: IL28B polymorphisms predict sustained viral response in HCV patients treated with pegylated interferon- $\alpha$  and ribavirin. *Aliment. Pharmacol. Ther.* 36, 91–103.

Ciccozzi, M., Equestre, M., Costantino, A., Marascio, N., Quirino, A., Presti, A.L., Cella, E., Bruni, R., Liberto, M.C., Foca, A., Pisani, G., Zehended, G., Ciccaglione, A.R., 2012. Hepatitis C virus genotype 4d in Southern Italy: reconstruction of its origin and spread by a phylodynamic analysis. *J. Med. Virol.* 84, 1613–1619.

De Bruijne, J., Schinkel, J., Prins, M., Koekkoek, S.M., Aronson, S.J., van Ballegooijen, M.W., Reesink, H.W., Molenkamp, R., van de Laar, T.J.W., 2009. Emergence of hepatitis C virus genotype 4: phylogenetic analysis reveals three distinct epidemiological profiles. *J. Clin. Microbiol.* 27, 3832–3838.

Djoko, C.F., Rimoin, A.W., Vidal, N., Tamoufe, U., Wolfe, N.D., Butel, C., LeBreton, M., Tshala, F.M., Kayembe, P.K., et al., 2011. High HIV type 1 group M pol diversity and low rate of antiretroviral resistance mutations among the uniformed services in Kinshasa, Democratic Republic of the Congo. *AIDS Res. Hum. Retrovir.* 27, 323–329.

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* 29, 1969–1973.

El-Zanaty, F., Way, A., 2009. *Egypt Demographic and Health Survey 2008—Final Report*. Report no. FR220. The DHS Program, Rockville, MD.



- Felsenstein, J., 1989. PHYLIP—phylogeny inference package (version 3.2). *Cladistics* 5, 164–166.
- Frank, C., Mohamed, M.K., Strickland, G.T., Lavanchy, D., Arthur, R.R., Magder, L.S., El Khoby, T., Abdel-Wahab, Y., Aly Ohn, E.S., Anwar, W., Sallam, I., 2000. The role of parenteral antischistosomal therapy in the spread of Hepatitis C Virus in Egypt. *Lancet* 355, 887–891.
- Gray, R.R., Parker, J., Lemey, P., Salemi, M., Katzourakis, A., Pybus, O.G., 2011. The mode and tempo of hepatitis C virus evolution within and among hosts. *BMC Evol. Biol.* 11, 131.
- Gray, R.R., Tanaka, Y., Takebe, Y., Magiorkinis, G., Buskell, Z., Seeff, L., Alter, H., Pybus, O.G., 2013. Evolutionary analysis of hepatitis C virus gene sequences from 1953. *Philos. Trans. R. Soc. Ser. B* 368, 20130168.
- Iles, J.C., Abby Harrison, G.L., Lyons, S., Djoko, C.F., Tamoufe, U., LeBreton, M., Schneider, B.S., Fair, J.N., Tshala, F.M., et al., 2013. Hepatitis C virus infections in the Democratic Republic of Congo exhibit a cohort effect. *Infect. Genet. Evol.* 19, 386–394.
- Jeannel, D., Fretz, C., Traore, Y., Kohdjo, N., Bigot, A., Gamy, E.P., Jourdan, G., Kourouma, K., Maertens, G., Fumoux, F., Fournel, J.J., Stuyver, L., 1998. Evidence for high genetic diversity and long-term endemicity of hepatitis C virus genotypes 1 and 2 in West Africa. *J. Med. Virol.* 55, 92–97.
- Kuiken, C., Yusim, K., Boykin, L., Richardson, R., 2005. The Los Alamos hepatitis C sequence database. *Bioinformatics* 21, 379–384.
- Lauer, G.M., Walker, B.D., 2001. Hepatitis C virus infection. *N. Engl. J. Med.* 345, 41–52.
- Laurent, C., Henzel, D., Mulanga-Kabeya, C., Maertens, G., Larouze, B., Delaporte, E., 2001. Seroepidemiological survey of hepatitis C virus among commercial sex workers and pregnant women in Kinshasa, Democratic Republic of Congo. *Int. J. Epidemiol.* 30, 872–877.
- Lemey, P., Rambaut, A., Drummond, A.J., Suchard, M.A., 2009. Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* 5, e1000520.
- Magiorkinis, G., Magiorkinis, E., Paraskevis, D., Ho, S.Y.W., Shapiro, B., Pybus, O.G., Allain, J.-P., Hatzakis, A., 2009. The global spread of hepatitis C virus 1a and 1b: a phylogenetic and phylogeographic analysis. *PLoS Med.* 6, e1000198.
- Markov, P.V., Pépin, J., Frost, E., Deslandes, S., Labbé, A.C., Pybus, O.G., 2009. Phylogeography and molecular epidemiology of hepatitis C virus genotype 2 in Africa. *J. Gen. Virol.* 90, 2086–2096.
- Markov, P.V., van de Laar, T.J., Thomas, X.V., Aronson, S.J., Weegink, C.J., van den Berk, G.E., Prins, M., Pybus, O.G., Schinkel, J., 2012. Colonial history and contemporary transmission shape the genetic diversity of hepatitis C genotype 2 in Amsterdam. *J. Virol.* 86, 7677–7687.
- Mays, T.M., 2010. *Historical Dictionary of Multinational Peacekeeping*. Scarecrow Press, Washington DC p. 321.
- Mellor, J., Holmes, E.C., Mellor, J., Jarvis, L.M., Yap, P.L., Holmes, E.C., Simmonds, P., Jarvis, L.M., Yap, P.L., Simmonds, P., 1995. Investigation of the pattern of hepatitis C virus sequence diversity in different geographical regions: implications for virus classification. The International HCV Collaborative Study Group. *J. Gen. Virol.* 76, 2493–2507.
- Mohd Hanafiah, K., Groeger, J., Flaxman, A.D., Wiersma, S.T., 2013. Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence. *Hepatology* 57, 1333–1342.
- Murphy, D.G., Chamberland, J., Dandavino, R., Sablon, E., 2007a. A new genotype of hepatitis C virus originating from Central Africa. *Hepatology* 46, 623A.
- Murphy, D.G., Willems, B., Deschenes, M., Hilzenrat, N., Mousseau, R., Sabbah, S., 2007b. Use of Sequence analysis of the NS5B region for routine genotyping of hepatitis C virus with reference to C/E1 and 5′ untranslated region sequences. *J. Clin. Microbiol.* 45, 1102–1112.
- Ndjomou, J., Pybus, O.G., Matz, B., 2003. Phylogenetic analysis of hepatitis C virus isolates indicates a unique pattern of endemic infection in Cameroon. *J. Gen. Virol.* 84, 2333–2341.
- Ndong-Atome, G., Makuwa, M., Njoum, R., Branger, M., Brun-Vézinet, F., Mahé, A., Rousset, D., Kazanji, M., 2008. Hepatitis C virus prevalence and genetic diversity among pregnant women in Gabon, central Africa. *BMC Infect. Dis.* 8, 82.
- Nerrienet, E., Pouillot, R., Lachenal, G., Njoum, R., Mfoupouendoun, J., Bilong, C., Mauciere, P., Pasquier, C., Ayouba, A., 2005. Hepatitis C virus infection in Cameroon: a cohort-effect. *J. Med. Virol.* 76, 208–214.
- Nicot, F., Legrand-Abravanel, F., Sandres-Saune, K., Boulestin, A., Dubois, M., Alric, L., Vinel, J.P., Pasquier, C., Izopet, J., 2005. Heterogeneity of hepatitis C virus genotype 4 strains circulating in south-western France. *J. Gen. Virol.* 86, 107–114.
- Nigel, T., 1991. *Foreign Volunteers of the Allied Forces, 1939–45*. Osprey, London p. 17.
- Njoum, R., Nerrienet, E., Dubois, M., Lachenal, G., Rousset, D., Vessière, A., Ayouba, A., Pasquier, C., Pouillot, R., 2007. The hepatitis C virus epidemic in Cameroon: genetic evidence for rapid transmission between 1920 and 1960. *Infect. Genet. Evol.* 7, 361–367.
- Njoum, R., Frost, E., Deslandes, S., Mamadou-Yaya, F., Labbé, A.-C., Pouillot, R., Mbélesso, P., Mbadingai, S., Rousset, D., Pépin, J., 2009. Predominance of hepatitis C virus genotype 4 infection and rapid transmission between 1935 and 1965 in the Central African Republic. *J. Gen. Virol.* 90, 2452–2456.
- Njoum, R., Caron, M., Besson, G., Ndong-Atome, G.-R., Makuwa, M., Pouillot, R., Nkoghe, D., Leroy, E., Kazanji, M., 2012. Phylogeography, risk factors and genetic history of hepatitis C virus in Gabon, central Africa. *PLoS One* 7, e42002.
- Pépin, J., Labbé, A.-C., 2008. Noble goals, unforeseen consequences: control of tropical diseases in colonial Central Africa and the iatrogenic transmission of blood-borne viruses. *Trop. Med. Int. Health* 13, 744–753.
- Pépin, J., Lavoie, M., Pybus, O.G., Pouillot, R., Foupouapouognigni, Y., Rousset, D., Labbé, A.-C., Njoum, R., 2010a. Risk factors for hepatitis C virus transmission in colonial Cameroon. *Clin. Infect. Dis.* 51, 768–776.
- Pépin, J., Labbé, A.C., Mamadou-Yaya, F., Mbélesso, P., Mbadingai, S., Deslandes, S., Locas, M.C., Frost, E., 2010b. Iatrogenic transmission of human T cell lymphotropic virus type 1 and hepatitis C virus through parenteral treatment and chemoprophylaxis of sleeping sickness in colonial Equatorial Africa. *Clin. Infect. Dis.* 51, 777–784.
- Pybus, O.G., Charleston, M.A., Gupta, S., Rambaut, A., Holmes, E.C., Harvey, P.H., 2001. The epidemic behavior of the hepatitis C virus. *Science* 292, 2323–2325.
- Pybus, O.G., Drummond, A.J., Nakano, T., Robertson, B.H., Rambaut, A., 2003. The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a s. *Mol. Biol. Evol.* 20, 381–387.
- Pybus, O.G., Cochrane, A., Holmes, E.C., Simmonds, P., 2005. The hepatitis C virus epidemic among injecting drug users. *Infect. Genet. Evol.* 5, 131–139.
- Pybus, O.G., Barnes, E., Taggart, R., Lemey, P., Markov, P.V., Rasachak, B., Syhavong, B., Phetsouvanah, R., Sheridan, I., Humphreys, I.S., Lu, L., Newton, P.N., Klennerman, P., 2009. Genetic history of hepatitis C virus in East Asia. *J. Virol.* 83, 1071–1082.
- Rambaut, A., Robertson, D.L., Pybus, O.G., Peeters, M., Holmes, E.C., 2001. Human immunodeficiency virus. Phylogeny and the origin of HIV-1. *Nature* 410, 1047–1048.
- Rose, R., Markov, P.V., Lam, T.T., Pybus, O.G., 2013. Viral evolution explains the associations among hepatitis C virus genotype, clinical outcomes, and human genetic variation. *Infect. Genet. Evol.* 20, 418–421.
- Salemi, M., Vandamme, A.-M., 2002. Hepatitis C virus evolutionary patterns studied through analysis of full-genome sequences. *J. Mol. Evol.* 54, 62–70.
- Shapiro, B., Rambaut, A., Drummond, A.J., 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol. Biol. Evol.* 23, 7–9.
- Sharp, C.P., Vermeulen, M., Nébié, Y., Djoko, C.F., LeBreton, M., Tamoufe, U., Rimoin, A.W., Kayembe, P.K., Carr, J.K., Servant-Delmas, A., Laperche, S., Harrison, G.L.A., Pybus, O.G., Delwart, E., Wolfe, N.D., Saville, A., Lefrère, J.-J., Simmonds, P., 2010. Epidemiology of human parvovirus 4 infection in sub-Saharan Africa. *Emerg. Infect. Dis.* 16, 1605–1607.
- Simmonds, P., Holmes, E.C., Cha, T.A., Chan, S.W., McOmish, F., Irvine, B., Beall, E., Yap, P.L., Kolberg, J., Urdea, M.S., 1993. Classification of hepatitis C virus into six major genotypes and a series of subtypes by phylogenetic analysis of the NS-5 region. *J. Gen. Virol.* 74, 2391–2399.
- Simmonds, P., 2004. Genetic diversity and evolution of hepatitis C virus—15 years on. *J. Gen. Virol.* 85, 3173–3188.
- Smith, D.B., Bukh, J., Kuiken, C., Muerhoff, A.S., Rice, C.M., Stapleton, J.T., Simmonds, P., 2014. Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource. *Hepatology* 59, 318–327.
- Strickland, G.T., 2006. Liver disease in Egypt: hepatitis C superseded schistosomiasis as a result of iatrogenic and biological factors. *Hepatology* 43, 915–922.
- Suret-Canale, J., 1971. *French Colonialism in tropical Africa 1900–1945*. Hurst and Company, London, pp. 462–484.
- Tibbs, C.J., Palmer, S.J., Coker, R., Clark, S.K., Parsons, G.M., Hojvat, S., Peterson, D., Banatvala, J.E., 1991. Prevalence of hepatitis C in tropical communities: the importance of confirmatory assays. *J. Med. Virol.* 34, 143–147.
- van Asten, L., Verhaest, I., Lamzira, S., Hernandez-Aguado, I., Zangerle, R., Boufassa, F., Rezza, G., Broers, B., Robertson, J.R., Brettell, R.P., McMenamin, J., Prins, M., Cochrane, A., Simmonds, P., Coutinho, R.A., 2004. Spread of hepatitis C virus among European injection drug users infected with HIV: a phylogenetic analysis. *J. Infect. Dis.* 189, 292–302.
- Vidal, N., Peeters, M., Mulanga-Kabeya, C., Nzilambi, N., Robertson, D., Ilunga, W., Sema, H., Tshimanga, K., Bongo, B., Delaporte, E., 2000. Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. *J. Virol.* 74, 10498–10507.
- Wertheim, J.O., Fournet, M., Kosakovsky Pond, S.L., 2012. Inconsistencies in estimating the age of HIV-1 subtypes due to heterotachy. *Mol. Biol. Evol.* 29, 451–456.
- Wertheim, J.O., Kosakovsky Pond, S.L., 2011. Purifying selection can obscure the ancient age of viral lineages. *Mol. Biol. Evol.* 28, 3355–3365.
- Zwickl, D.J., 2006. *Genetic Algorithm Approaches For The Phylogenetic Analysis Of Large Biological Sequence Datasets Under The Maximum Likelihood Criterion*. The University of Texas at Austin, Austin, Texas, USA.